PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-245067
(43)Date of publication of application : 30.08.2002

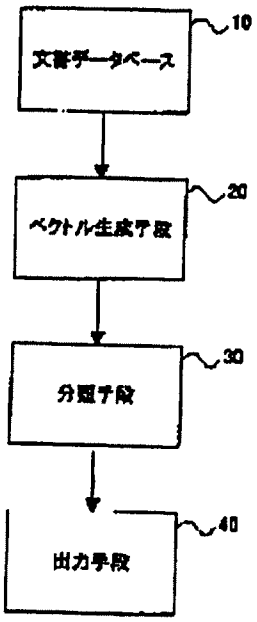
(51)Int.Cl. G06F 17/30

(21)Application number : 2001-037163
(22)Date of filing : 14.02.2001

(71)Applicant : MITSUBISHI ELECTRIC CORP
(72)Inventor : KONAKA HIROYOSHI
TSUDAKA SHINICHIRO
KOBUNE RYUICHI
ARITA HIDEKAZU

(54) INFORMATION RETRIEVAL UNIT

(57)Abstract:
PROBLEM TO BE SOLVED: To obtain an information retrieval unit for calculating a similarity degree which reflects relation between keywords and improving precision in classification or retrieval.
SOLUTION: The unit is provided with a document database 10 storing multiple kinds of document data, a vector generating means 20 for generating the feature vector of the keyword concerning each kind of document data, a classifying means 30 for calculating the similarity degree between the feature vectors and classifying document data and an output means 40 for outputting the classification result of document data. The vector generating means 20 analyzes the respective kinds of document data, extracts the keywords and relation between the keywords and generates the feature vector based on the appearance frequency of the both.



Cited Reference ②

JP Laid-Open Patent Publication No. 2002-245067

CLAIMS

[Claim(s)]
[Claim 1]A document data base which stores two or more document data.
A vector generating means which generates a feature vector to each above-mentioned document data.
A sorting means which calculates similarity between the above-mentioned feature vectors, and classifies each above-mentioned document data.
An output means which outputs a classification result of the above-mentioned document data.
It is the information retrieval device provided with the above, and the above-mentioned vector generating means analyzes each above-mentioned document data respectively, extracts a relation between keywords, and generates the above-mentioned feature vector based on both frequency of occurrence of these.

[Claim 2]A document data base which stores two or more document data.
A search formula input means which inputs a search formula.
A vector generating means which generates a feature vector to each above-mentioned document data and the above-mentioned search formula.
A similarity calculation means to calculate similarity between a feature vector to the above-mentioned search formula, and a feature vector to each above-mentioned document data.
An output means which outputs document data which has the above-mentioned high feature vector of similarity.
It is the information retrieval device provided with the above, and the above-mentioned vector generating means analyzes respectively each above-mentioned document data and a search formula, extracts a relation between keywords, and generates the above-mentioned feature vector based on these frequencies of occurrence.

[Claim 3]The information retrieval device according to claim 1 or 2, wherein a relation of dependency is used for the above-mentioned vector generating means as a relation between the above-mentioned keywords.
[Claim 4]The information retrieval device according to claim 1 or 2 using that the above-mentioned vector generating means has a near distance between keywords as a relation between the above-mentioned keywords.
[Claim 5]The above-mentioned vector generating means instead of the frequency of occurrence of a relation between keywords containing a keyword contained in a keyword group belonging to the same category, or it, The information retrieval device according to any one of claims 1 to 4 using what added those frequencies of occurrence, respectively as the frequency of occurrence of a relation between keywords containing a keyword or it representing the category.
[Claim 6]The information retrieval device according to any one of claims 1 to 5, wherein the above-mentioned vector generating means generates a feature vector based on weighting which a user specifies to the frequency of occurrence of a relation between keywords.

[Translation done.]

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention]This invention relates especially document data to the information retrieval device which carries out classification and search automatically about a classification and search of the electronized document data.
[0002]
[Description of the Prior Art]About a classification and search of the electronized document data, the information retrieval device shown in the former, for example, JP,11-110395,A, is proposed. In the information retrieval device proposed here, the synonymous frequency of occurrence is summarized, a feature vector is generated, the similarity between feature vectors is calculated, and each document data is classified. Attaching weighting to two or more words which have a synonymous relation, respectively is also proposed by this JP,11-110395,A.
[0003]While summarizing the synonymous frequency of occurrence in JP,10-198691,A, and generating a feature vector is indicated, the adjoining word pair in a document data base and the synonym pair are registered, for example, and using for calculation of a feature vector is indicated.
[0004]
[Problem(s) to be Solved by the Invention]In the conventional information retrieval device of such composition, therefore the dependency of words and phrases, etc. had not carried out the classification or search reflecting the relation between keywords, a high-precision classification or search was not able to be carried out.
[0005]This invention was made in order to solve above SUBJECT, and it makes possible similarity calculation not only reflecting a keyword but the relation between keywords, and an object of an invention is to obtain the information retrieval device which can improve the accuracy of a classification or search.
[0006]
[Means for Solving the Problem]A document data base with which an information retrieval device concerning this invention stores two or more document data, In an information retrieval device which has a vector generating means which generates a feature vector to each document data, a sorting means which calculates similarity between feature vectors and classifies each document data, and an output means which outputs a classification result of document data, A vector generating means analyzes each document data respectively, extracts a relation between keywords, and generates a feature vector based on both frequency of occurrence of these.
[0007]A document data base with which an information retrieval device concerning this invention stores two or more document data, A search formula input means which inputs a search formula, and a vector generating means which generates a feature vector to each document data and search formula, In an information retrieval device which has a similarity calculation means to calculate similarity between a feature vector to a search formula, and a feature vector to each document data, and an output means which outputs document data which has a high feature vector of similarity, A vector generating means analyzes each document data and a search formula respectively, extracts a relation between keywords, and generates a feature vector

based on these frequencies of occurrence.
[0008]A relation of dependency is used for a vector generating means as a relation between keywords.
[0009]It is used for a vector generating means as a relation between keywords that distance between keywords is near.
[0010]Instead of the frequency of occurrence of a relation between keywords in which a vector generating means contains a keyword contained in a keyword group belonging to the same category, or it, What added those frequencies of occurrence, respectively is used as the frequency of occurrence of a relation between keywords containing a keyword or it representing the category.
[0011]A vector generating means generates a feature vector based on weighting which a user specifies to the frequency of occurrence of a relation between keywords.
[0012]
[Embodiment of the Invention]Embodiment 1. drawing 1 is a block diagram showing the example of composition of the information retrieval device about the classification of this invention. In a figure, the document data base 10 stores two or more document data. Each document data stored in the document data base 10 has text data at least.
[0013]The vector generating means 20 generates a feature vector to each document data. That is, conduct a morphological analysis etc. to the text data of each document data, perform unnecessary word processing etc. if needed, and a keyword is extracted, and the relation between keywords is extracted.
[0014]Next, when R relations between K keywords are extracted from the document data base 10 whole which consists of N documents, the feature vector Vi of each document i ($1 <= i <= N$) is expressed with the vector of a K+R dimension, for example. When the index of the relation between keywords is expressed with j ($1 <= j <= K+R$), according to the tf-idf method, the ingredient Vij of each dimension j of the feature vector Vi can be computed by the following formula, for example.
[0015]$Vij = TFij * \log(N/DFj)$
[0016]Here, TFij is [be / it / under / document i / setting] the number of times in which the relation between the keywords corresponding to j ingredient appears, and DFj is the number of times in which the relation between the keywords corresponding to j ingredient appears in N whole sentence in the letter of the document data base 10. Thus, a feature vector is generated.
[0017]In the sorting means 30, the similarity between documents is calculated and a document is clustered using the result. The similarity between documents is calculable with the cosine value of the angle between each feature vector of two or more sentence document computed as mentioned above, for example. By using the similarity calculated about clustering using the feature vector generated as mentioned above to similarity calculation required for clustering algorithms, such as K method of averaging. Clustering not only reflecting the keyword used by the conventional clustering but the relation between keywords is attained.
[0018]The classified result can be outputted by the output means 40.
[0019]Here, what have the relation of the dependency obtained as a result of syntax analysis and the distance between keywords near as an example of the related extraction between the keywords at the time of generating a feature vector to each document data can be considered by the vector generating means 20.
[0020]First, the sentence "C A carries out B" is considered about the relation of dependency, for example. In this sentence, the relation of the dependency of "C A carrying out" and "C Carrying out B" exists. Although it may identify including these to a rank, a rank being disregarded, and "A->C", "B->C", or a direction also being disregarded, and considering it as "A&C" and "B&C" (it considers that "A->C" and "C->A" are the same) is also considered. The appearance frequency of such dependency will be used as said TFij concerning the relation between the keywords in this case, or DFj.
[0021]As an example of the distance in the case of on the other hand using what has a near distance between keywords as a relation between keywords, the number of characters between keywords, the number of morphemes, the number of clauses, the number of sentences, the

number of paragraphs, etc. can be considered, for example. The case where a direction is considered also in this case may not be considered. As said TFij concerning the relation between the keywords in this case, or DFj, appearance frequency when this distance is smaller than a user designated value will be used, for example.
[0022]It is also possible to perform the classification reflecting the category of keywords, such as synonymous-words-related. Namely, if it explains taking the case of the relation of dependency, supposing the keyword a0 and a1 belong to the category A and a0 and a1 have the relation between b and dependency, The keyword a0, the dimension about a1, "a0", "a0->b", "a1", and "a1->b" can be summarized to "A" and "A->b", and a feature vector can also be generated.
[0023]If neither of each dimension of the feature vector used as the comparison object in the case of a classification is a non-zero ingredient (coincidence), it will not contribute to similarity. However, generally, since coincidence of the relation between keywords becomes low probable rather than coincidence of a mere keyword, compared with a keyword, there is a tendency for the contribution to the similarity of the relation between keywords to become low. Then, balance of contribution to both similarity evaluation can be aimed at by making dignity of the dimension about the relation between keywords larger than the dimension of a keyword.
[0024]It is also possible for a user to select a keyword, to enlarge dignity of the dimension of the relation between the keywords in which the dimension of the keyword and its keyword are contained, and to perform the classification which thought as important the keyword which a user observes.
[0025]Embodiment 2. drawing 2 is a block diagram showing the example of composition of the information retrieval device about search which are other examples of this invention, and the document data base 10 and the output means 40 have the same function as drawing 1.
[0026]Have the search formula input means 50 and the function to input a search condition as a search formula (the text expressing a search formula may be sufficient.) the vector generating means 20, About the whole sentence document stored in the document data base 10, generate a feature vector from the frequency of occurrence of the relation between keywords according to the feature vector formula described by Embodiment 1, for example, and. It has a function which generates a feature vector also from the inputted search formula using the feature vector generation method described by Embodiment 1, and the same method. However, when generating a feature vector from a search formula, i of said Vij shall not show the document i and shall show a search formula. In this case, TFij is usually set to 1.
[0027]The search means 60 the similarity between the feature vector generated from the search formula, and the feature vector about the whole sentence document in the stored document data base 10. It calculates in the method described by the means of the classification of Embodiment 1, and a similar way, and similarity ranking evaluation of the document in a document data base is performed using the result.
[0028]The result which carried out ranking attachment can be outputted by the output means 40.
[0029]By the vector generating means 20, it is possible to use the relation which a keyword requires, the number of things with a near distance between keywords, etc. the same [with having stated by Embodiment 1] as an example of the related extraction between the keywords at the time of generating a feature vector to each document data.
[0030]It is also possible to perform search reflecting the category of keywords, such as synonymous-words-related. How to the category of the relation between keywords to collect is the same as that of illustration of Embodiment 1.
[0031]By what dignity of the dimension of the relation between specific keywords is enlarged for. It is also possible to aim at balance of the dignity of each dimension of the relation between keywords, or to perform search which thought as important the relation between the keywords which a user observes like illustration by Embodiment 1.
[0032]
[Effect of the Invention]The document data base with which the information retrieval device concerning this invention stores two or more document data, In the information retrieval device

which has a vector generating means which generates a feature vector to each document data, a sorting means which calculates the similarity between feature vectors and classifies each document data, and an output means which outputs the classification result of document data, A vector generating means analyzes each document data respectively, extracts the relation between keywords, and generates a feature vector based on both frequency of occurrence of these. Therefore, in a classification of document data, the similarity calculation not only reflecting the keyword of each document data but the relation between keywords becomes possible, and accuracy improves.

[0033]The document data base with which the information retrieval device concerning this invention stores two or more document data, The search formula input means which inputs a search formula, and the vector generating means which generates a feature vector to each document data and search formula, In the information retrieval device which has a similarity calculation means to calculate the similarity between the feature vector to a search formula, and the feature vector to each document data, and an output means which outputs the document data which has a high feature vector of similarity, A vector generating means analyzes each document data and a search formula respectively, extracts the relation between keywords, and generates a feature vector based on these frequencies of occurrence. Therefore, in search of the document data near a search formula, the similarity calculation not only reflecting the keyword which appears in a search formula and each document data but the relation between keywords becomes possible, and the accuracy of search improves.

[0034]The relation of dependency is used for a vector generating means as a relation between keywords. Therefore, in a classification of document data or search of the document data near a search formula, similarity calculation reflecting the relation of dependency is performed and the accuracy of a classification or search improves compared with the conventional method only using a keyword.

[0035]It is used for a vector generating means as a relation between keywords that the distance between keywords is near. Therefore, in a classification of document data or search of the document data near a search formula, similarity calculation reflecting the distance between keywords is performed, and the accuracy of a classification or search improves compared with the conventional method only using a keyword.

[0036]Instead of the frequency of occurrence of the relation between the keywords in which a vector generating means contains the keyword contained in the keyword group belonging to the same category, or it, What added those frequencies of occurrence, respectively is used as the frequency of occurrence of the relation between the keywords containing the keyword or it representing the category. Therefore, in a classification of document data or search of the document data near a search formula, similarity calculation which summarized the relation between the keywords which do not need distinguishing for a user can be performed. As a result, it becomes possible to attain the increase in efficiency of highly precise classification and search.

[0037]A vector generating means generates a feature vector based on weighting which a user specifies to the frequency of occurrence of the relation between keywords. Therefore, in a classification of document data or search of the document data near a search formula, a user's intention is reflected and similarity calculation which thought as important or made light of the relation between specific keywords can be performed. As a result, highly precise-ization of a classification and search in the form which reflected a user's intention better is attained.

[Translation done.]

TECHNICAL FIELD

[Field of the Invention]This invention relates especially document data to the information retrieval device which carries out classification and search automatically about a classification and search of the electronized document data.

PRIOR ART

[Description of the Prior Art]About a classification and search of the electronized document data, the information retrieval device shown in the former, for example, JP,11-110395,A, is proposed. In the information retrieval device proposed here, the synonymous frequency of occurrence is summarized, a feature vector is generated, the similarity between feature vectors is calculated, and each document data is classified. Attaching weighting to two or more words which have a synonymous relation, respectively is also proposed by this JP,11-110395,A.
[0003]While summarizing the synonymous frequency of occurrence in JP,10-198691,A, and generating a feature vector is indicated, the adjoining word pair in a document data base and the synonym pair are registered, for example, and using for calculation of a feature vector is indicated.

[Translation done.]

EFFECT OF THE INVENTION

[Effect of the Invention]The document data base with which the information retrieval device concerning this invention stores two or more document data, In the information retrieval device which has a vector generating means which generates a feature vector to each document data, a sorting means which calculates the similarity between feature vectors and classifies each document data, and an output means which outputs the classification result of document data, A vector generating means analyzes each document data respectively, extracts the relation between keywords, and generates a feature vector based on both frequency of occurrence of these. Therefore, in a classification of document data, the similarity calculation not only reflecting the keyword of each document data but the relation between keywords becomes possible, and accuracy improves.
[0033]The document data base with which the information retrieval device concerning this invention stores two or more document data, The search formula input means which inputs a search formula, and the vector generating means which generates a feature vector to each document data and search formula, In the information retrieval device which has a similarity calculation means to calculate the similarity between the feature vector to a search formula, and the feature vector to each document data, and an output means which outputs the document data which has a high feature vector of similarity, A vector generating means analyzes each document data and a search formula respectively, extracts the relation between keywords, and generates a feature vector based on these frequencies of occurrence. Therefore, in search of the document data near a search formula, the similarity calculation not only reflecting the keyword which appears in a search formula and each document data but the relation between keywords becomes possible, and the accuracy of search improves.
[0034]The relation of dependency is used for a vector generating means as a relation between keywords. Therefore, in a classification of document data or search of the document data near a search formula, similarity calculation reflecting the relation of dependency is performed and the accuracy of a classification or search improves compared with the conventional method only using a keyword.
[0035]It is used for a vector generating means as a relation between keywords that the distance between keywords is near. Therefore, in a classification of document data or search of the document data near a search formula, similarity calculation reflecting the distance between keywords is performed, and the accuracy of a classification or search improves compared with the conventional method only using a keyword.
[0036]Instead of the frequency of occurrence of the relation between the keywords in which a vector generating means contains the keyword contained in the keyword group belonging to the same category, or it, What added those frequencies of occurrence, respectively is used as the frequency of occurrence of the relation between the keywords containing the keyword or it representing the category. Therefore, in a classification of document data or search of the document data near a search formula, similarity calculation which summarized the relation between the keywords which do not need distinguishing for a user can be performed. As a result, it becomes possible to attain the increase in efficiency of highly precise classification and search.

[0037]A vector generating means generates a feature vector based on weighting which a user specifies to the frequency of occurrence of the relation between keywords. Therefore, in a classification of document data or search of the document data near a search formula, a user's intention is reflected and similarity calculation which thought as important or made light of the relation between specific keywords can be performed. As a result, highly precise-ization of a classification and search in the form which reflected a user's intention better is attained.

[Translation done.]

TECHNICAL PROBLEM

[Problem(s) to be Solved by the Invention]In the conventional information retrieval device of such composition, therefore the dependency of words and phrases, etc. had not carried out the classification or search reflecting the relation between keywords, a high-precision classification or search was not able to be carried out.
[0005]This invention was made in order to solve above SUBJECT, and it makes possible similarity calculation not only reflecting a keyword but the relation between keywords, and an object of an invention is to obtain the information retrieval device which can improve the accuracy of a classification or search.

[Translation done.]

MEANS

[Means for Solving the Problem]A document data base with which an information retrieval device concerning this invention stores two or more document data, In an information retrieval device which has a vector generating means which generates a feature vector to each document data, a sorting means which calculates similarity between feature vectors and classifies each document data, and an output means which outputs a classification result of document data, A vector generating means analyzes each document data respectively, extracts a relation between keywords, and generates a feature vector based on both frequency of occurrence of these.
[0007]A document data base with which an information retrieval device concerning this invention stores two or more document data, A search formula input means which inputs a search formula, and a vector generating means which generates a feature vector to each document data and search formula, In an information retrieval device which has a similarity calculation means to calculate similarity between a feature vector to a search formula, and a feature vector to each document data, and an output means which outputs document data which has a high feature vector of similarity, A vector generating means analyzes each document data and a search formula respectively, extracts a relation between keywords, and generates a feature vector based on these frequencies of occurrence.
[0008]A relation of dependency is used for a vector generating means as a relation between keywords.
[0009]It is used for a vector generating means as a relation between keywords that distance between keywords is near.
[0010]Instead of the frequency of occurrence of a relation between keywords in which a vector generating means contains a keyword contained in a keyword group belonging to the same category, or it, What added those frequencies of occurrence, respectively is used as the frequency of occurrence of a relation between keywords containing a keyword or it representing the category.
[0011]A vector generating means generates a feature vector based on weighting which a user specifies to the frequency of occurrence of a relation between keywords.
[0012]
[Embodiment of the Invention]Embodiment 1. drawing 1 is a block diagram showing the example of composition of the information retrieval device about the classification of this invention. In a figure, the document data base 10 stores two or more document data. Each document data stored in the document data base 10 has text data at least.
[0013]The vector generating means 20 generates a feature vector to each document data. That is, conduct a morphological analysis etc. to the text data of each document data, perform unnecessary word processing etc. if needed, and a keyword is extracted, and the relation between keywords is extracted.
[0014]Next, when R relations between K keywords are extracted from the document data base 10 whole which consists of N documents, the feature vector Vi of each document i ($1 \leq i \leq N$) is expressed with the vector of a K+R dimension, for example. When the index of the relation between keywords is expressed with j ($1 \leq j \leq K+R$), according to the tf-idf method, the ingredient Vij of each dimension j of the feature vector Vi can be computed by the following

formula, for example.
[0015]Vij=TFij*log(N/DFj)
[0016]Here, TFij is [be / it / under / document i / setting] the number of times in which the relation between the keywords corresponding to j ingredient appears, and DFj is the number of times in which the relation between the keywords corresponding to j ingredient appears in N whole sentence in the letter of the document data base 10. Thus, a feature vector is generated.
[0017]In the sorting means 30, the similarity between documents is calculated and a document is clustered using the result. The similarity between documents is calculable with the cosine value of the angle between each feature vector of two or more sentence document computed as mentioned above, for example. By using the similarity calculated about clustering using the feature vector generated as mentioned above to similarity calculation required for clustering algorithms, such as K method of averaging, Clustering not only reflecting the keyword used by the conventional clustering but the relation between keywords is attained.
[0018]The classified result can be outputted by the output means 40.
[0019]Here, what have the relation of the dependency obtained as a result of syntax analysis and the distance between keywords near as an example of the related extraction between the keywords at the time of generating a feature vector to each document data can be considered by the vector generating means 20.
[0020]First, the sentence "C A carries out B" is considered about the relation of dependency, for example. In this sentence, the relation of the dependency of "C A carrying out" and "C Carrying out B" exists. Although it may identify including these to a rank, a rank being disregarded, and "A->C", "B->C", or a direction also being disregarded, and considering it as "A&C" and "B&C" (it considers that "A->C" and "C->A" are the same) is also considered. The appearance frequency of such dependency will be used as said TFij concerning the relation between the keywords in this case, or DFj.
[0021]As an example of the distance in the case of on the other hand using what has a near distance between keywords as a relation between keywords, the number of characters between keywords, the number of morphemes, the number of clauses, the number of sentences, the number of paragraphs, etc. can be considered, for example. The case where a direction is considered also in this case may not be considered. As said TFij concerning the relation between the keywords in this case, or DFj, appearance frequency when this distance is smaller than a user designated value will be used, for example.
[0022]It is also possible to perform the classification reflecting the category of keywords, such as synonymous-words-related. Namely, if it explains taking the case of the relation of dependency, supposing the keyword a0 and a1 belong to the category A and a0 and a1 have the relation between b and dependency, The keyword a0, the dimension about a1, "a0", "a0->b", "a1", and "a1->b" can be summarized to "A" and "A->b", and a feature vector can also be generated.
[0023]If neither of each dimension of the feature vector used as the comparison object in the case of a classification is a non-zero ingredient (coincidence), it will not contribute to similarity. However, generally, since coincidence of the relation between keywords becomes low probable rather than coincidence of a mere keyword, compared with a keyword, there is a tendency for the contribution to the similarity of the relation between keywords to become low. Then, balance of contribution to both similarity evaluation can be aimed at by making dignity of the dimension about the relation between keywords larger than the dimension of a keyword.
[0024]It is also possible for a user to select a keyword, to enlarge dignity of the dimension of the relation between the keywords in which the dimension of the keyword and its keyword are contained, and to perform the classification which thought as important the keyword which a user observes.
[0025]Embodiment 2. drawing 2 is a block diagram showing the example of composition of the information retrieval device about search which are other examples of this invention, and the document data base 10 and the output means 40 have the same function as drawing 1.
[0026]Have the search formula input means 50 and the function to input a search condition as a search formula (the text expressing a search formula may be sufficient.) the vector generating

means 20. About the whole sentence document stored in the document data base 10, generate a feature vector from the frequency of occurrence of the relation between keywords according to the feature vector formula described by Embodiment 1, for example, and. It has a function which generates a feature vector also from the inputted search formula using the feature vector generation method described by Embodiment 1, and the same method. However, when generating a feature vector from a search formula, i of said Vij shall not show the document i and shall show a search formula. In this case, TFij is usually set to 1.

[0027]The search means 60 the similarity between the feature vector generated from the search formula, and the feature vector about the whole sentence document in the stored document data base 10, It calculates in the method described by the means of the classification of Embodiment 1, and a similar way, and similarity ranking evaluation of the document in a document data base is performed using the result.

[0028]The result which carried out ranking attachment can be outputted by the output means 40.

[0029]By the vector generating means 20, it is possible to use the relation which a keyword requires, the number of things with a near distance between keywords, etc. the same [with having stated by Embodiment 1] as an example of the related extraction between the keywords at the time of generating a feature vector to each document data.

[0030]It is also possible to perform search reflecting the category of keywords, such as synonymous-words-related. How to the category of the relation between keywords to collect is the same as that of illustration of Embodiment 1.

[0031]By what dignity of the dimension of the relation between specific keywords is enlarged for. It is also possible to aim at balance of the dignity of each dimension of the relation between keywords, or to perform search which thought as important the relation between the keywords which a user observes like illustration by Embodiment 1.

[Translation done.]

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1]It is a block diagram showing the information retrieval device relevant to the classification of this invention.
[Drawing 2]It is a block diagram showing the information retrieval device relevant to search of this invention.
[Description of Notations]
10 A document data base, 20 vector generating means, and 30 A sorting means, 40 output means, and 50 A search formula input means and 60 Similarity calculation means (search means).

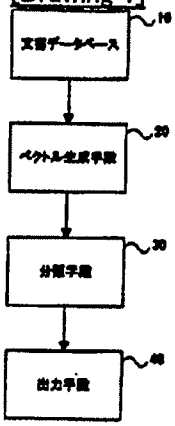
[Translation done.]

* NOTICES *

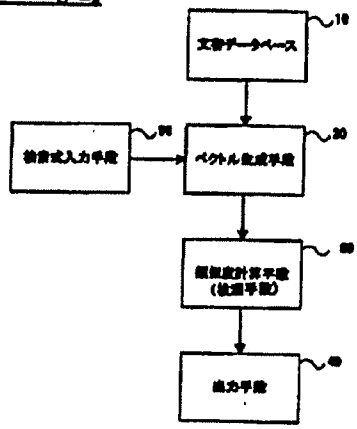
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

[Drawing 1]



[Drawing 2]



[Translation done.]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報(A)　(11)特許出願公開番号
特開2002－245067
(P2002－245067A)
(43)公開日　平成14年８月30日(2002.8.30)

(51)Int.Cl.'	識別記号	FI	テーマコード'(参考)
G06F 17/30	210	G06F 17/30	210D 5B075
	170		170A
	340		340B
	350		350C

審査請求　未請求　請求項の数6　OL　(全５頁)

(21)出願番号　特願2001－37163(P2001－37163)

(22)出願日　平成13年２月14日(2001.2.14)

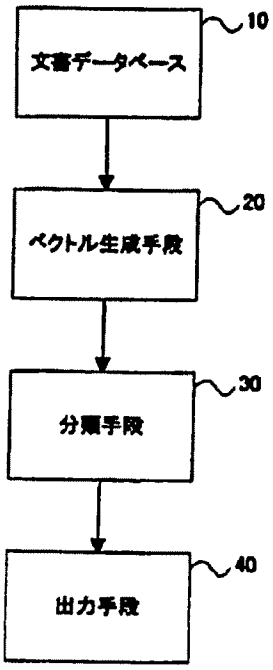
(出願人による申告）国等の委託研究の成果に係る特許出願（平成12年度、通商産業省「軽水炉等改良技術確証試験等（発電設備診断システム開発）の委託研究、産業活力再生特別措置法第30条の適用を受けるもの）

(71)出願人　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目２番３号
(72)発明者　小中　裕喜
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内
(72)発明者　津高　新一郎
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内
(74)代理人　100057874
弁理士　曾我　道照　(外４名)

最終頁に続く

(54)【発明の名称】　情報検索装置

(57)【要約】
【課題】　キーワード間の関係を反映した類似度計算を可能とし、分類あるいは検索の精度を向上することができる情報検索装置を得る。
【解決手段】　複数の文書データを格納する文書データベース１０と、各々の文書データに対しキーワードの特徴ベクトルを生成するベクトル生成手段２０と、特徴ベクトル間の類似度を計算して各文書データを分類する分類手段３０と、文書データの分類結果を出力する出力手段４０とを有する情報検索装置において、ベクトル生成手段２０は、各文書データを各々解析してキーワード及びキーワード間の関係を抽出し、これら両方の出現頻度に基づいて特徴ベクトルを生成する。



【特許請求の範囲】
【請求項１】　複数の文書データを格納する文書データベースと、
各々の上記文書データに対し特徴ベクトルを生成するベクトル生成手段と、
上記特徴ベクトル間の類似度を計算して上記各文書データを分類する分類手段と、
上記文書データの分類結果を出力する出力手段とを有する情報検索装置において、
上記ベクトル生成手段は、上記各文書データを各々解析してキーワード及びキーワード間の関係を抽出し、これら両方の出現頻度に基づいて上記特徴ベクトルを生成することを特徴とする情報検索装置。
【請求項２】　複数の文書データを格納する文書データベースと、
検索式を入力する検索式入力手段と、
各々の上記文書データ及び上記検索式に対し特徴ベクトルを生成するベクトル生成手段と、
上記検索式に対する特徴ベクトルと各々の上記文書データに対する特徴ベクトル間の類似度を計算する類似度計算手段と、
類似度の高い上記特徴ベクトルを有する文書データを出力する出力手段とを有する情報検索装置において、
上記ベクトル生成手段は、上記各文書データ及び検索式を各々解析してキーワード及びキーワード間の関係を抽出し、これらの出現頻度に基づいて上記特徴ベクトルを生成することを特徴とする情報検索装置。
【請求項３】　上記ベクトル生成手段は、上記キーワード間の関係として係り受けの関係を用いることを特徴とする請求項１または２に記載の情報検索装置。
【請求項４】　上記ベクトル生成手段は、上記キーワード間の関係としてキーワード間の距離が近いことを用いることを特徴とする請求項１または２に記載の情報検索装置。
【請求項５】　上記ベクトル生成手段は、同一カテゴリに属するキーワード群に含まれるキーワードもしくはそれを含むキーワード間の関係の出現頻度の代わりに、そのカテゴリを代表するキーワードもしくはそれを含むキーワード間の関係の出現頻度としてそれらの出現頻度をそれぞれ加算したものを用いることを特徴とする請求項１から４のいずれかに記載の情報検索装置。
【請求項６】　上記ベクトル生成手段は、キーワード及びキーワード間の関係の出現頻度に対し、利用者が指定する重みづけに基づいて特徴ベクトルを生成することを特徴とする請求項１から５のいずれかに記載の情報検索装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】この発明は、電子化された文書データの分類・検索に関し、特に文書データを自動的に分類・検索をする情報検索装置に関するものである。
【０００２】
【従来の技術】電子化された文書データの分類・検索に関しては、従来、例えば特開平１１－１１０３９５号公報に示されている情報検索装置が提案されている。ここに提案されている情報検索装置においては、類義語の出現頻度をまとめて特徴ベクトルを生成し、特徴ベクトル間の類似度を計算して各文書データを分類する。また、この特開平１１－１１０３９５号公報には、類義語の関係にある複数の単語にそれぞれ重み付けを付けることも提案されている。
【０００３】さらに、例えば特開平１０－１９８６９１号公報においては、類義語の出現頻度をまとめて特徴ベクトルを生成することが開示されているとともに、文書データベース中の隣接した単語対、類義語対を登録しておき、特徴ベクトルの計算に用いることが開示されている。
【０００４】
【発明が解決しようとする課題】このような構成の従来の情報検索装置においては、語句の係り受け等、キーワード間の関係を反映した分類あるいは検索をしていない、そのため、精度の高い分類あるいは検索をすることが出来なかった。
【０００５】この発明は、上述のような課題を解決するためになされたもので、キーワードだけでなくキーワード間の関係をも反映した類似度計算を可能とし、分類あるいは検索の精度を向上することができる情報検索装置を得ることを目的とする。
【０００６】
【課題を解決するための手段】この発明に係る情報検索装置は、複数の文書データを格納する文書データベースと、各々の文書データに対し特徴ベクトルを生成するベクトル生成手段と、特徴ベクトル間の類似度を計算して各文書データを分類する分類手段と、文書データの分類結果を出力する出力手段とを有する情報検索装置において、ベクトル生成手段は、各文書データを各々解析してキーワード及びキーワード間の関係を抽出し、これら両方の出現頻度に基づいて特徴ベクトルを生成する。
【０００７】また、この発明に係る情報検索装置は、複数の文書データを格納する文書データベースと、検索式を入力する検索式入力手段と、各々の文書データ及び検索式に対し特徴ベクトルを生成するベクトル生成手段と、検索式に対する特徴ベクトルと各々の文書データに対する特徴ベクトル間の類似度を計算する類似度計算手段と、類似度の高い特徴ベクトルを有する文書データを出力する出力手段とを有する情報検索装置において、ベクトル生成手段は、各文書データ及び検索式を各々解析してキーワード及びキーワード間の関係を抽出し、これらの出現頻度に基づいて特徴ベクトルを生成する。
【０００８】また、ベクトル生成手段は、キーワード間

の関係として係り受けの関係を用いる。

【0009】また、ベクトル生成手段は、キーワード間の関係としてキーワード間の距離が近いことを用いる。

【0010】また、ベクトル生成手段は、同一カテゴリに属するキーワード群に含まれるキーワードもしくはそれを含むキーワード間の関係の出現頻度の代わりに、そのカテゴリを代表するキーワードもしくはそれを含むキーワード間の関係の出現頻度としてそれらの出現頻度をそれぞれ加算したものを用いる。

【0011】さらに、ベクトル生成手段は、キーワード及びキーワード間の関係の出現頻度に対し、利用者が指定する重みづけに基づいて特徴ベクトルを生成する。

【0012】

【発明の実施の形態】実施の形態1．図1は、この発明の分類に関する情報検索装置の構成例を示すブロック図である。図において文書データベース10は複数の文書データを格納する。文書データベース10に格納された各文書データは少なくともテキストデータを有している。

【0013】ベクトル生成手段20は、各文書データに対して特徴ベクトルを生成する。すなわち、各文書データのテキストデータに対して形態素解析などを行い、必要に応じて不要語処理等を行ってキーワードを抽出すると共に、キーワード間の関係を抽出する。

【0014】次に、N個の文書からなる文書データベース10全体から、キーワードK個、キーワード間の関係R個が抽出されたとき、各文書i（$1 \leq i \leq N$）の特徴ベクトルViは、たとえば、K+R次元のベクトルで表される。キーワード若しくはキーワード間の関係のインデックスをj（$1 \leq j \leq K+R$）で表すとき、特徴ベクトルViの各次元jの成分Vijは、たとえば、tf・idf法によると、次の式で算出できる。

【0015】$Vij = TFij * log(N/DFj)$

【0016】ここで、TFijは、文書i中において、j成分に対応するキーワード若しくはキーワード間の関係が現れる回数であり、また、DFjは、文書データベース10のN個の全文書中において、j成分に対応するキーワード若しくはキーワード間の関係が現れる回数である。このようにして、特徴ベクトルが生成される。

【0017】分類手段30では、文書間の類似度を計算すると共に、その結果を使って、文書のクラスタリングを行う。文書間の類似度は、たとえば上記のように算出した複数文書の各特徴ベクトル間の角度のコサイン値で計算できる。クラスタリングについては、K平均法などのクラスタリングアルゴリズムに必要な類似度計算に、上述のように生成した特徴ベクトルを使って計算した類似度を用いることにより、従来のクラスタリングで用いられていたキーワードだけでなく、キーワード間の関係をも反映したクラスタリングが可能となる。

【0018】分類した結果は出力手段40により出力することができる。

【0019】ここで、ベクトル生成手段20で、各文書データに対して特徴ベクトルを生成する際のキーワード間の関係抽出の具体例としては、構文解析の結果として得られる係り受けの関係やキーワード間の距離の近いものなどが考えられる。

【0020】まず、係り受けの関係について、たとえば、「AがBをCする」という文を考える。この文においては、「AがCする」「BをCする」という係り受けの関係が存在する。これらを格まで含めて識別してもよいが、格を無視して、「A→C」、「B→C」あるいは、方向も無視して、「A&C」、「B&C」（「A→C」と「C→A」を同じと見なす）とすることも考えられる。この場合のキーワード間の関係に係る、前記TFij、若しくはDFjとしてはこのような係り受けの出現回数を使用することになる。

【0021】一方、キーワード間の関係として、キーワード間の距離の近いものを用いる場合の距離の具体例としては、たとえばキーワード間の文字数、形態素数、文節数、文数、段落数等が考えられる。この場合も方向を考える場合と考えない場合が存在する。この場合のキーワード間の関係に係る、前記TFij、若しくはDFjとしては、例えば、この距離がユーザー指定値より小さい場合の出現回数を使用することになる。

【0022】さらに、同義語関係などのキーワードのカテゴリを反映した分類を行うことも可能である。すなわち、係り受けの関係を例にとって説明すれば、キーワードa0、a1がカテゴリAに属しており、a0、a1はbと係り受けの関係があるとすれば、キーワードa0、a1に関する次元、「a0」、「a0→b」、「a1」、「a1→b」を、「A」、「A→b」に、まとめて、特徴ベクトルを生成することも出来る。

【0023】なお、分類の際の比較対象となる特徴ベクトルの各次元は、共に非ゼロ成分（共起）でなければ類似度には寄与しない。しかしながら、一般に、キーワード間の関係の共起は、単なるキーワードの共起よりも確率的に低くなるため、キーワードに比べて、キーワード間の関係の類似度への寄与が低くなってしまう傾向がある。そこで、キーワード間の関係に関する次元の重みをキーワードの次元よりも大きくすることにより、両者の類似度評価への寄与のバランスを図ることができる。

【0024】また、ユーザーがキーワードを選定して、そのキーワードの次元やそのキーワードが含まれるキーワード間の関係の次元の重みを大きくして、ユーザーが注目するキーワードを重視した分類を行うことも可能である。

【0025】実施の形態2．図2は、この発明の他の実施例である、検索に関する情報検索装置の構成例を示すブロック図であり、文書データベース10、出力手段40は図1と同じ機能を有する。

【００２６】検索式入力手段５０は、検索条件を検索式（検索式を表現する文章でも良い。）として入力する機能を有し、ベクトル生成手段２０は、文書データベース１０に格納されている全文書について、キーワード及びキーワード間の関係の出現頻度から、例えば、実施の形態１で述べた特徴ベクトル計算式にしたがって特徴ベクトルを生成すると共に、入力した検索式からも、実施の形態１で述べた特徴ベクトル生成方法と同様の方法を用いて、特徴ベクトルを生成する機能を有する。（ただし、検索式から特徴ベクトルを生成する場合は、前記ＶＩｊのｉは文書ｉを示すものではなく、検索式を示すものとする。この場合、ＴＦＩｊは通常１になる。）

【００２７】検索手段６０は、検索式から生成した特徴ベクトルと、格納された文書データベース１０中の全文書についての特徴ベクトル間の類似度を、実施の形態１の分類の手段で述べた方法と同様な方法で計算し、その結果を使って、文書データベース中の文書の類似度ランキング評価を行う。

【００２８】ランキング付けした結果は出力手段４０により出力することができる。

【００２９】なお、ベクトル生成手段２０では、各文書データに対して特徴ベクトルを生成する際のキーワード間の関係抽出の具体例として、実施の形態１で述べたことと同様に、キーワードの係り付けの関係やキーワード間の距離の近いもの数などを利用することが考えられる。

【００３０】さらに、同義語関係などのキーワードのカテゴリを反映した検索を行うことも可能である。キーワード若しくはキーワード間の関係のカテゴリーへのまとめ方は、実施の形態１の例示と同様である。

【００３１】また、特定のキーワード若しくはキーワード間の関係の次元の重みを大きくすることで、実施の形態１での例示と同様に、キーワード／キーワード間の関係のそれぞれの次元の重みのバランスを図ったり、ユーザーが注目するキーワード若しくはキーワード間の関係を重視した検索を行うことも可能である。

【００３２】

【発明の効果】この発明に係る情報検索装置は、複数の文書データを格納する文書データベースと、各々の文書データに対し特徴ベクトルを生成するベクトル生成手段と、特徴ベクトル間の類似度を計算して各文書データを分類する分類手段と、文書データの分類結果を出力する出力手段とを有する情報検索装置において、ベクトル生成手段は、各文書データを各々解析してキーワード及びキーワード間の関係を抽出し、これら両方の出現頻度に基づいて特徴ベクトルを生成する。そのため、文書データの分類において、各文書データのキーワードだけでなく、キーワード間の関係をも反映した類似度計算が可能となり精度が向上する。

【００３３】また、この発明に係る情報検索装置は、複数の文書データを格納する文書データベースと、検索式を入力する検索式入力手段と、各々の文書データ及び検索式に対し特徴ベクトルを生成するベクトル生成手段と、検索式に対する特徴ベクトルと各々の文書データに対する特徴ベクトル間の類似度を計算する類似度計算手段と、類似度の高い特徴ベクトルを有する文書データを出力する出力手段とを有する情報検索装置において、ベクトル生成手段は、各文書データ及び検索式を各々解析してキーワード及びキーワード間の関係を抽出し、これらの出現頻度に基づいて特徴ベクトルを生成する。そのため、検索式に近い文書データの検索において、検索式および各文書データに出現するキーワードだけでなく、キーワード間の関係をも反映した類似度計算が可能となり検索の精度が向上する。

【００３４】また、ベクトル生成手段は、キーワード間の関係として係り受けの関係を用いる。そのため、文書データの分類や検索式に近い文書データの検索において、係り受けの関係を反映した類似度計算が行われ、キーワードのみを用いた従来の方式に比べ分類や検索の精度が向上する。

【００３５】また、ベクトル生成手段は、キーワード間の関係としてキーワード間の距離が近いことを用いる。そのため、文書データの分類や検索式に近い文書データの検索において、キーワード間の距離を反映した類似度計算が行われ、キーワードのみを用いた従来の方式に比べ分類や検索の精度が向上する。

【００３６】また、ベクトル生成手段は、同一カテゴリに属するキーワード群に含まれるキーワードもしくはそれを含むキーワード間の関係の出現頻度の代わりに、そのカテゴリを代表するキーワードもしくはそれを含むキーワード間の関係の出現頻度としてそれらの出現頻度をそれぞれ加算したものを用いる。そのため、文書データの分類や検索式に近い文書データの検索において、利用者にとって区別することが不要なキーワードあるいはキーワード間の関係をまとめた類似度計算を行うことができる。その結果、高精度の分類・検索の効率化を図ることが可能となる。

【００３７】さらに、ベクトル生成手段は、キーワード及びキーワード間の関係の出現頻度に対し、利用者が指定する重みづけに基づいて特徴ベクトルを生成する。そのため、文書データの分類や検索式に近い文書データの検索において、利用者の意図を反映して、特定のキーワードあるいはキーワード間の関係を重視あるいは軽視した類似度計算を行うことができる。その結果、利用者の意図をより良く反映した形での分類・検索の高精度化が可能となる。

【図面の簡単な説明】

【図１】　この発明の分類に関連する情報検索装置を示すブロック図である。

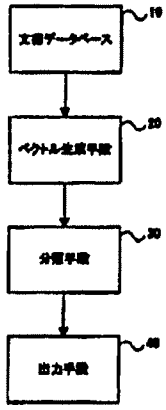
【図２】　この発明の検索に関連する情報検索装置を示

すブロック図である。

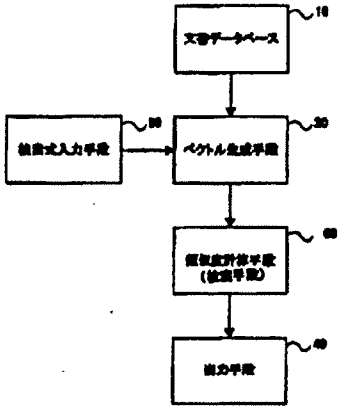
【符号の説明】

１０　文書データベース、２０　ベクトル生成手段、３＊０　分類手段、４０出力手段、５０　検索式入力手段、６０　類似度計算手段（検索手段）。

【図１】



【図２】



フロントページの続き

(72)発明者　小船　隆一
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内

(72)発明者　有田　英一
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内

Ｆターム(参考)　5B075 ND03 NK02 NR12 PP23 PR04 PR06 QM08

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-245067
(43)Date of publication of application : 30.08.2002

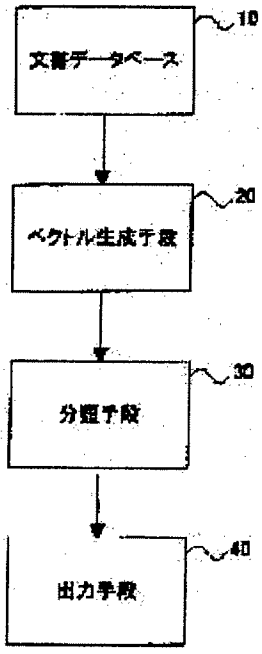
(51)Int.Cl. G06F 17/30

(21)Application number : 2001-037163
(22)Date of filing : 14.02.2001

(71)Applicant : MITSUBISHI ELECTRIC CORP
(72)Inventor : KONAKA HIROYOSHI
TSUDAKA SHINICHIRO
KOBUNE RYUICHI
ARITA HIDEKAZU

(54) INFORMATION RETRIEVAL UNIT

(57)Abstract:
PROBLEM TO BE SOLVED: To obtain an information retrieval unit for calculating a similarity degree which reflects relation between keywords and improving precision in classification or retrieval.
SOLUTION: The unit is provided with a document database 10 storing multiple kinds of document data, a vector generating means 20 for generating the feature vector of the keyword concerning each kind of document data, a classifying means 30 for calculating the similarity degree between the feature vectors and classifying document data and an output means 40 for outputting the classification result of document data. The vector generating means 20 analyzes the respective kinds of document data, extracts the keywords and relation between the keywords and generates the feature vector based on the appearance frequency of the both.



CLAIMS

[Claim(s)]
[Claim 1]A document data base which stores two or more document data.
A vector generating means which generates a feature vector to each above-mentioned document data.
A sorting means which calculates similarity between the above-mentioned feature vectors, and classifies each above-mentioned document data.
An output means which outputs a classification result of the above-mentioned document data.
It is the information retrieval device provided with the above, and the above-mentioned vector generating means analyzes each above-mentioned document data respectively, extracts a relation between keywords, and generates the above-mentioned feature vector based on both frequency of occurrence of these.

[Claim 2]A document data base which stores two or more document data.
A search formula input means which inputs a search formula.
A vector generating means which generates a feature vector to each above-mentioned document data and the above-mentioned search formula.
A similarity calculation means to calculate similarity between a feature vector to the above-mentioned search formula, and a feature vector to each above-mentioned document data.
An output means which outputs document data which has the above-mentioned high feature vector of similarity.
It is the information retrieval device provided with the above, and the above-mentioned vector generating means analyzes respectively each above-mentioned document data and a search formula, extracts a relation between keywords, and generates the above-mentioned feature vector based on these frequencies of occurrence.

[Claim 3]The information retrieval device according to claim 1 or 2, wherein a relation of dependency is used for the above-mentioned vector generating means as a relation between the above-mentioned keywords.
[Claim 4]The information retrieval device according to claim 1 or 2 using that the above-mentioned vector generating means has a near distance between keywords as a relation between the above-mentioned keywords.
[Claim 5]The above-mentioned vector generating means instead of the frequency of occurrence of a relation between keywords containing a keyword contained in a keyword group belonging to the same category, or it, The information retrieval device according to any one of claims 1 to 4 using what added those frequencies of occurrence, respectively as the frequency of occurrence of a relation between keywords containing a keyword or it representing the category.
[Claim 6]The information retrieval device according to any one of claims 1 to 5, wherein the above-mentioned vector generating means generates a feature vector based on weighting which a user specifies to the frequency of occurrence of a relation between keywords.

[Translation done.]

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Field of the Invention]This invention relates especially document data to the information retrieval device which carries out classification and search automatically about a classification and search of the electronized document data.
[0002]
[Description of the Prior Art]About a classification and search of the electronized document data, the information retrieval device shown in the former, for example, JP,11-110395,A, is proposed. In the information retrieval device proposed here, the synonymous frequency of occurrence is summarized, a feature vector is generated, the similarity between feature vectors is calculated, and each document data is classified. Attaching weighting to two or more words which have a synonymous relation, respectively is also proposed by this JP,11-110395,A.
[0003]While summarizing the synonymous frequency of occurrence in JP,10-198691,A, and generating a feature vector is indicated, the adjoining word pair in a document data base and the synonym pair are registered, for example, and using for calculation of a feature vector is indicated.
[0004]
[Problem(s) to be Solved by the Invention]In the conventional information retrieval device of such composition, therefore the dependency of words and phrases, etc. had not carried out the classification or search reflecting the relation between keywords, a high-precision classification or search was not able to be carried out.
[0005]This invention was made in order to solve above SUBJECT, and it makes possible similarity calculation not only reflecting a keyword but the relation between keywords, and an object of an invention is to obtain the information retrieval device which can improve the accuracy of a classification or search.
[0006]
[Means for Solving the Problem]A document data base with which an information retrieval device concerning this invention stores two or more document data, In an information retrieval device which has a vector generating means which generates a feature vector to each document data, a sorting means which calculates similarity between feature vectors and classifies each document data, and an output means which outputs a classification result of document data, A vector generating means analyzes each document data respectively, extracts a relation between keywords, and generates a feature vector based on both frequency of occurrence of these.
[0007]A document data base with which an information retrieval device concerning this invention stores two or more document data, A search formula input means which inputs a search formula, and a vector generating means which generates a feature vector to each document data and search formula, In an information retrieval device which has a similarity calculation means to calculate similarity between a feature vector to a search formula, and a feature vector to each document data, and an output means which outputs document data which has a high feature vector of similarity, A vector generating means analyzes each document data and a search formula respectively, extracts a relation between keywords, and generates a feature vector

based on these frequencies of occurrence.
[0008]A relation of dependency is used for a vector generating means as a relation between keywords.
[0009]It is used for a vector generating means as a relation between keywords that distance between keywords is near.
[0010]Instead of the frequency of occurrence of a relation between keywords in which a vector generating means contains a keyword contained in a keyword group belonging to the same category, or it, What added those frequencies of occurrence, respectively is used as the frequency of occurrence of a relation between keywords containing a keyword or it representing the category.
[0011]A vector generating means generates a feature vector based on weighting which a user specifies to the frequency of occurrence of a relation between keywords.
[0012]
[Embodiment of the Invention]Embodiment 1. drawing 1 is a block diagram showing the example of composition of the information retrieval device about the classification of this invention. In a figure, the document data base 10 stores two or more document data. Each document data stored in the document data base 10 has text data at least.
[0013]The vector generating means 20 generates a feature vector to each document data. That is, conduct a morphological analysis etc. to the text data of each document data, perform unnecessary word processing etc. if needed, and a keyword is extracted, and the relation between keywords is extracted.
[0014]Next, when R relations between K keywords are extracted from the document data base 10 whole which consists of N documents, the feature vector Vi of each document i (1 <=i<=N) is expressed with the vector of a K+R dimension, for example. When the index of the relation between keywords is expressed with j (1 <=j<=K+R), according to the tf-idf method, the ingredient Vij of each dimension j of the feature vector Vi can be computed by the following formula, for example.
[0015]Vij=TFij*log(N/DFj)
[0016]Here, TFij is [be / it / under / document i / setting] the number of times in which the relation between the keywords corresponding to j ingredient appears, and DFj is the number of times in which the relation between the keywords corresponding to j ingredient appears in N whole sentence in the letter of the document data base 10. Thus, a feature vector is generated.
[0017]In the sorting means 30, the similarity between documents is calculated and a document is clustered using the result. The similarity between documents is calculable with the cosine value of the angle between each feature vector of two or more sentence document computed as mentioned above, for example. By using the similarity calculated about clustering using the feature vector generated as mentioned above to similarity calculation required for clustering algorithms, such as K method of averaging, Clustering not only reflecting the keyword used by the conventional clustering but the relation between keywords is attained.
[0018]The classified result can be outputted by the output means 40.
[0019]Here, what have the relation of the dependency obtained as a result of syntax analysis and the distance between keywords near as an example of the related extraction between the keywords at the time of generating a feature vector to each document data can be considered by the vector generating means 20.
[0020]First, the sentence "C A carries out B" is considered about the relation of dependency, for example. In this sentence, the relation of the dependency of "C A carrying out" and "C Carrying out B" exists. Although it may identify including these to a rank, a rank being disregarded, and "A->C", "B->C", or a direction also being disregarded, and considering it as "A&C" and "B&C" (it considers that "A->C" and "C->A" are the same) is also considered. The appearance frequency of such dependency will be used as said TFij concerning the relation between the keywords in this case, or DFj.
[0021]As an example of the distance in the case of on the other hand using what has a near distance between keywords as a relation between keywords, the number of characters between keywords, the number of morphemes, the number of clauses, the number of sentences, the

number of paragraphs, etc. can be considered, for example. The case where a direction is considered also in this case may not be considered. As said TFij concerning the relation between the keywords in this case, or DFj, appearance frequency when this distance is smaller than a user designated value will be used, for example.
[0022]It is also possible to perform the classification reflecting the category of keywords, such as synonymous-words-related. Namely, if it explains taking the case of the relation of dependency, supposing the keyword a0 and a1 belong to the category A and a0 and a1 have the relation between b and dependency, The keyword a0, the dimension about a1, "a0", "a0->b", "a1", and "a1->b" can be summarized to "A" and "A->b", and a feature vector can also be generated.
[0023]If neither of each dimension of the feature vector used as the comparison object in the case of a classification is a non-zero ingredient (coincidence), it will not contribute to similarity. However, generally, since coincidence of the relation between keywords becomes low probable rather than coincidence of a mere keyword, compared with a keyword, there is a tendency for the contribution to the similarity of the relation between keywords to become low. Then, balance of contribution to both similarity evaluation can be aimed at by making dignity of the dimension about the relation between keywords larger than the dimension of a keyword.
[0024]It is also possible for a user to select a keyword, to enlarge dignity of the dimension of the relation between the keywords in which the dimension of the keyword and its keyword are contained, and to perform the classification which thought as important the keyword which a user observes.
[0025]Embodiment 2. drawing 2 is a block diagram showing the example of composition of the information retrieval device about search which are other examples of this invention, and the document data base 10 and the output means 40 have the same function as drawing 1.
[0026]Have the search formula input means 50 and the function to input a search condition as a search formula (the text expressing a search formula may be sufficient.) the vector generating means 20, About the whole sentence document stored in the document data base 10, generate a feature vector from the frequency of occurrence of the relation between keywords according to the feature vector formula described by Embodiment 1, for example, and. It has a function which generates a feature vector also from the inputted search formula using the feature vector generation method described by Embodiment 1, and the same method. However, when generating a feature vector from a search formula, i of said Vij shall not show the document i and shall show a search formula. In this case, TFij is usually set to 1.
[0027]The search means 60 the similarity between the feature vector generated from the search formula, and the feature vector about the whole sentence document in the stored document data base 10, It calculates in the method described by the means of the classification of Embodiment 1, and a similar way, and similarity ranking evaluation of the document in a document data base is performed using the result.
[0028]The result which carried out ranking attachment can be outputted by the output means 40.
[0029]By the vector generating means 20, it is possible to use the relation which a keyword requires, the number of things with a near distance between keywords, etc. the same [with having stated by Embodiment 1] as an example of the related extraction between the keywords at the time of generating a feature vector to each document data.
[0030]It is also possible to perform search reflecting the category of keywords, such as synonymous-words-related. How to the category of the relation between keywords to collect is the same as that of illustration of Embodiment 1.
[0031]By what dignity of the dimension of the relation between specific keywords is enlarged for. It is also possible to aim at balance of the dignity of each dimension of the relation between keywords, or to perform search which thought as important the relation between the keywords which a user observes like illustration by Embodiment 1.
[0032]
[Effect of the Invention]The document data base with which the information retrieval device concerning this invention stores two or more document data, In the information retrieval device

which has a vector generating means which generates a feature vector to each document data, a sorting means which calculates the similarity between feature vectors and classifies each document data, and an output means which outputs the classification result of document data, A vector generating means analyzes each document data respectively, extracts the relation between keywords, and generates a feature vector based on both frequency of occurrence of these. Therefore, in a classification of document data, the similarity calculation not only reflecting the keyword of each document data but the relation between keywords becomes possible, and accuracy improves.

[0033]The document data base with which the information retrieval device concerning this invention stores two or more document data, The search formula input means which inputs a search formula, and the vector generating means which generates a feature vector to each document data and search formula, In the information retrieval device which has a similarity calculation means to calculate the similarity between the feature vector to a search formula, and the feature vector to each document data, and an output means which outputs the document data which has a high feature vector of similarity, A vector generating means analyzes each document data and a search formula respectively, extracts the relation between keywords, and generates a feature vector based on these frequencies of occurrence. Therefore, in search of the document data near a search formula, the similarity calculation not only reflecting the keyword which appears in a search formula and each document data but the relation between keywords becomes possible, and the accuracy of search improves.

[0034]The relation of dependency is used for a vector generating means as a relation between keywords. Therefore, in a classification of document data or search of the document data near a search formula, similarity calculation reflecting the relation of dependency is performed and the accuracy of a classification or search improves compared with the conventional method only using a keyword.

[0035]It is used for a vector generating means as a relation between keywords that the distance between keywords is near. Therefore, in a classification of document data or search of the document data near a search formula, similarity calculation reflecting the distance between keywords is performed, and the accuracy of a classification or search improves compared with the conventional method only using a keyword.

[0036]Instead of the frequency of occurrence of the relation between the keywords in which a vector generating means contains the keyword contained in the keyword group belonging to the same category, or it, What added those frequencies of occurrence, respectively is used as the frequency of occurrence of the relation between the keywords containing the keyword or it representing the category. Therefore, in a classification of document data or search of the document data near a search formula, similarity calculation which summarized the relation between the keywords which do not need distinguishing for a user can be performed. As a result, it becomes possible to attain the increase in efficiency of highly precise classification and search.

[0037]A vector generating means generates a feature vector based on weighting which a user specifies to the frequency of occurrence of the relation between keywords. Therefore, in a classification of document data or search of the document data near a search formula, a user's intention is reflected and similarity calculation which thought as important or made light of the relation between specific keywords can be performed. As a result, highly precise-ization of a classification and search in the form which reflected a user's intention better is attained.

[Translation done.]

[0037]A vector generating means generates a feature vector based on weighting which a user specifies to the frequency of occurrence of the relation between keywords. Therefore, in a classification of document data or search of the document data near a search formula, a user's intention is reflected and similarity calculation which thought as important or made light of the relation between specific keywords can be performed. As a result, highly precise-ization of a classification and search in the form which reflected a user's intention better is attained.

[Translation done.]

TECHNICAL FIELD

[Field of the Invention]This invention relates especially document data to the information retrieval device which carries out classification and search automatically about a classification and search of the electronized document data.

[Translation done.]

TECHNICAL PROBLEM

[Problem(s) to be Solved by the Invention]In the conventional information retrieval device of such composition, therefore the dependency of words and phrases, etc. had not carried out the classification or search reflecting the relation between keywords, a high-precision classification or search was not able to be carried out.
[0005]This invention was made in order to solve above SUBJECT, and it makes possible similarity calculation not only reflecting a keyword but the relation between keywords, and an object of an invention is to obtain the information retrieval device which can improve the accuracy of a classification or search.

[Translation done.]

MEANS

[Means for Solving the Problem]A document data base with which an information retrieval device concerning this invention stores two or more document data, In an information retrieval device which has a vector generating means which generates a feature vector to each document data, a sorting means which calculates similarity between feature vectors and classifies each document data, and an output means which outputs a classification result of document data, A vector generating means analyzes each document data respectively, extracts a relation between keywords, and generates a feature vector based on both frequency of occurrence of these.
[0007]A document data base with which an information retrieval device concerning this invention stores two or more document data, A search formula input means which inputs a search formula, and a vector generating means which generates a feature vector to each document data and search formula, In an information retrieval device which has a similarity calculation means to calculate similarity between a feature vector to a search formula, and a feature vector to each document data, and an output means which outputs document data which has a high feature vector of similarity, A vector generating means analyzes each document data and a search formula respectively, extracts a relation between keywords, and generates a feature vector based on these frequencies of occurrence.
[0008]A relation of dependency is used for a vector generating means as a relation between keywords.
[0009]It is used for a vector generating means as a relation between keywords that distance between keywords is near.
[0010]Instead of the frequency of occurrence of a relation between keywords in which a vector generating means contains a keyword contained in a keyword group belonging to the same category, or it, What added those frequencies of occurrence, respectively is used as the frequency of occurrence of a relation between keywords containing a keyword or it representing the category.
[0011]A vector generating means generates a feature vector based on weighting which a user specifies to the frequency of occurrence of a relation between keywords.
[0012]
[Embodiment of the Invention]Embodiment 1. drawing 1 is a block diagram showing the example of composition of the information retrieval device about the classification of this invention. In a figure, the document data base 10 stores two or more document data. Each document data stored in the document data base 10 has text data at least.
[0013]The vector generating means 20 generates a feature vector to each document data. That is, conduct a morphological analysis etc. to the text data of each document data, perform unnecessary word processing etc. if needed, and a keyword is extracted, and the relation between keywords is extracted.
[0014]Next, when R relations between K keywords are extracted from the document data base 10 whole which consists of N documents, the feature vector Vi of each document i (1 <=i<=N) is expressed with the vector of a K+R dimension, for example. When the index of the relation between keywords is expressed with j (1 <=j<=K+R), according to the tf-idf method, the ingredient Vij of each dimension j of the feature vector Vi can be computed by the following

formula, for example.
[0015]Vij=TFij*log(N/DFj)
[0016]Here, TFij is [be / it / under / document i / setting] the number of times in which the relation between the keywords corresponding to j ingredient appears, and DFj is the number of times in which the relation between the keywords corresponding to j ingredient appears in N whole sentence in the letter of the document data base 10. Thus, a feature vector is generated.
[0017]In the sorting means 30, the similarity between documents is calculated and a document is clustered using the result. The similarity between documents is calculable with the cosine value of the angle between each feature vector of two or more sentence document computed as mentioned above, for example. By using the similarity calculated about clustering using the feature vector generated as mentioned above to similarity calculation required for clustering algorithms, such as K method of averaging, Clustering not only reflecting the keyword used by the conventional clustering but the relation between keywords is attained.
[0018]The classified result can be outputted by the output means 40.
[0019]Here, what have the relation of the dependency obtained as a result of syntax analysis and the distance between keywords near as an example of the related extraction between the keywords at the time of generating a feature vector to each document data can be considered by the vector generating means 20.
[0020]First, the sentence "C A carries out B" is considered about the relation of dependency, for example. In this sentence, the relation of the dependency of "C A carrying out" and "C Carrying out B" exists. Although it may identify including these to a rank, a rank being disregarded, and "A->C", "B->C", or a direction also being disregarded, and considering it as "A&C" and "B&C" (it considers that "A->C" and "C->A" are the same) is also considered. The appearance frequency of such dependency will be used as said TFij concerning the relation between the keywords in this case, or DFj.
[0021]As an example of the distance in the case of on the other hand using what has a near distance between keywords as a relation between keywords, the number of characters between keywords, the number of morphemes, the number of clauses, the number of sentences, the number of paragraphs, etc. can be considered, for example. The case where a direction is considered also in this case may not be considered. As said TFij concerning the relation between the keywords in this case, or DFj, appearance frequency when this distance is smaller than a user designated value will be used, for example.
[0022]It is also possible to perform the classification reflecting the category of keywords, such as synonymous-words-related. Namely, if it explains taking the case of the relation of dependency, supposing the keyword a0 and a1 belong to the category A and a0 and a1 have the relation between b and dependency, The keyword a0, the dimension about a1, "a0", "a0->b", "a1", and "a1->b" can be summarized to "A" and "A->b", and a feature vector can also be generated.
[0023]If neither of each dimension of the feature vector used as the comparison object in the case of a classification is a non-zero ingredient (coincidence), it will not contribute to similarity. However, generally, since coincidence of the relation between keywords becomes low probable rather than coincidence of a mere keyword, compared with a keyword, there is a tendency for the contribution to the similarity of the relation between keywords to become low. Then, balance of contribution to both similarity evaluation can be aimed at by making dignity of the dimension about the relation between keywords larger than the dimension of a keyword.
[0024]It is also possible for a user to select a keyword, to enlarge dignity of the dimension of the relation between the keywords in which the dimension of the keyword and its keyword are contained, and to perform the classification which thought as important the keyword which a user observes.
[0025]Embodiment 2. drawing 2 is a block diagram showing the example of composition of the information retrieval device about search which are other examples of this invention, and the document data base 10 and the output means 40 have the same function as drawing 1.
[0026]Have the search formula input means 50 and the function to input a search condition as a search formula (the text expressing a search formula may be sufficient.) the vector generating

means 20, About the whole sentence document stored in the document data base 10, generate a feature vector from the frequency of occurrence of the relation between keywords according to the feature vector formula described by Embodiment 1, for example, and. It has a function which generates a feature vector also from the inputted search formula using the feature vector generation method described by Embodiment 1, and the same method. However, when generating a feature vector from a search formula, i of said Vij shall not show the document i and shall show a search formula. In this case, TFij is usually set to 1.

[0027]The search means 60 the similarity between the feature vector generated from the search formula, and the feature vector about the whole sentence document in the stored document data base 10, It calculates in the method described by the means of the classification of Embodiment 1, and a similar way, and similarity ranking evaluation of the document in a document data base is performed using the result.

[0028]The result which carried out ranking attachment can be outputted by the output means 40.

[0029]By the vector generating means 20, it is possible to use the relation which a keyword requires, the number of things with a near distance between keywords, etc. the same [with having stated by Embodiment 1] as an example of the related extraction between the keywords at the time of generating a feature vector to each document data.

[0030]It is also possible to perform search reflecting the category of keywords, such as synonymous-words-related. How to the category of the relation between keywords to collect is the same as that of illustration of Embodiment 1.

[0031]By what dignity of the dimension of the relation between specific keywords is enlarged for. It is also possible to aim at balance of the dignity of each dimension of the relation between keywords, or to perform search which thought as important the relation between the keywords which a user observes like illustration by Embodiment 1.

[Translation done.]

EFFECT OF THE INVENTION

[Effect of the Invention]The document data base with which the information retrieval device concerning this invention stores two or more document data, In the information retrieval device which has a vector generating means which generates a feature vector to each document data, a sorting means which calculates the similarity between feature vectors and classifies each document data, and an output means which outputs the classification result of document data, A vector generating means analyzes each document data respectively, extracts the relation between keywords, and generates a feature vector based on both frequency of occurrence of these. Therefore, in a classification of document data, the similarity calculation not only reflecting the keyword of each document data but the relation between keywords becomes possible, and accuracy improves.

[0033]The document data base with which the information retrieval device concerning this invention stores two or more document data, The search formula input means which inputs a search formula, and the vector generating means which generates a feature vector to each document data and search formula, In the information retrieval device which has a similarity calculation means to calculate the similarity between the feature vector to a search formula, and the feature vector to each document data, and an output means which outputs the document data which has a high feature vector of similarity, A vector generating means analyzes each document data and a search formula respectively, extracts the relation between keywords, and generates a feature vector based on these frequencies of occurrence. Therefore, in search of the document data near a search formula, the similarity calculation not only reflecting the keyword which appears in a search formula and each document data but the relation between keywords becomes possible, and the accuracy of search improves.

[0034]The relation of dependency is used for a vector generating means as a relation between keywords. Therefore, in a classification of document data or search of the document data near a search formula, similarity calculation reflecting the relation of dependency is performed and the accuracy of a classification or search improves compared with the conventional method only using a keyword.

[0035]It is used for a vector generating means as a relation between keywords that the distance between keywords is near. Therefore, in a classification of document data or search of the document data near a search formula, similarity calculation reflecting the distance between keywords is performed, and the accuracy of a classification or search improves compared with the conventional method only using a keyword.

[0036]Instead of the frequency of occurrence of the relation between the keywords in which a vector generating means contains the keyword contained in the keyword group belonging to the same category, or it, What added those frequencies of occurrence, respectively is used as the frequency of occurrence of the relation between the keywords containing the keyword or it representing the category. Therefore, in a classification of document data or search of the document data near a search formula, similarity calculation which summarized the relation between the keywords which do not need distinguishing for a user can be performed. As a result, it becomes possible to attain the increase in efficiency of highly precise classification and search.

PRIOR ART

[Description of the Prior Art]About a classification and search of the electronized document data, the information retrieval device shown in the former, for example, JP,11-110395,A, is proposed. In the information retrieval device proposed here, the synonymous frequency of occurrence is summarized, a feature vector is generated, the similarity between feature vectors is calculated, and each document data is classified. Attaching weighting to two or more words which have a synonymous relation, respectively is also proposed by this JP,11-110395,A.
[0003]While summarizing the synonymous frequency of occurrence in JP,10-198691,A, and generating a feature vector is indicated, the adjoining word pair in a document data base and the synonym pair are registered, for example, and using for calculation of a feature vector is indicated.

[Translation done.]

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1]It is a block diagram showing the information retrieval device relevant to the classification of this invention.
[Drawing 2]It is a block diagram showing the information retrieval device relevant to search of this invention.
[Description of Notations]
10 A document data base, 20 vector generating means, and 30 A sorting means, 40 output means, and 50 A search formula input means and 60 Similarity calculation means (search means).

[Translation done.]

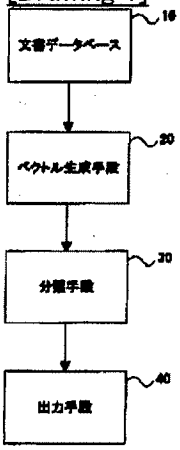
* NOTICES *

JPO and INPIT are not responsible for any damages caused by the use of this translation.

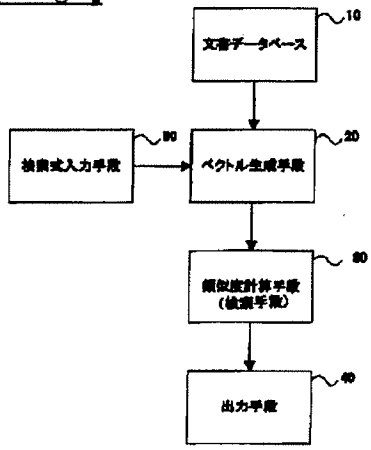
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

[Drawing 1]



[Drawing 2]



[Translation done.]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2002－245067
（P2002－245067A）
(43)公開日　平成14年８月30日(2002.8.30)

(51)Int.Cl.'	識別記号	ＦＩ	テーマコード'（参考）
Ｇ０６Ｆ　17/30	２１０	Ｇ０６Ｆ　17/30	２１０Ｄ　５Ｂ０７５
	１７０		１７０Ａ
	３４０		３４０Ｂ
	３５０		３５０Ｃ

審査請求　未請求　請求項の数６　ＯＬ　（全５頁）

(21)出願番号　特願2001－37163(P2001－37163)

(22)出願日　平成13年２月14日(2001.2.14)

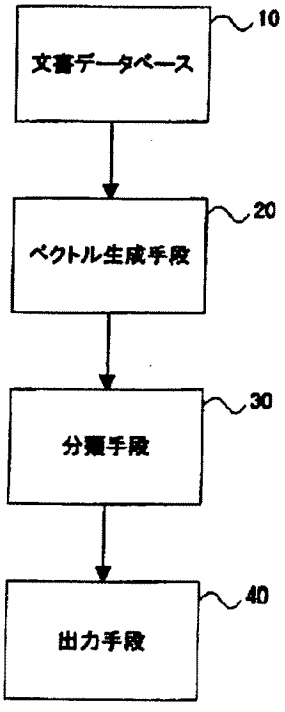
（出願人による申告）国等の委託研究の成果に係る特許出願（平成12年度、通商産業省「軽水炉等改良技術確証試験等（発電設備診断シテム開発）の委託研究、産業活力再生特別措置法第30条の適用を受けるもの）

(71)出願人　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目２番３号
(72)発明者　小中　裕喜
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内
(72)発明者　津高　新一郎
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内
(74)代理人　100057874
弁理士　曾我　道照　（外４名）

最終頁に続く

(54)【発明の名称】　情報検索装置

(57)【要約】
【課題】　キーワード間の関係を反映した類似度計算を可能とし、分類あるいは検索の精度を向上することができる情報検索装置を得る。
【解決手段】　複数の文書データを格納する文書データベース１０と、各々の文書データに対しキーワードの特徴ベクトルを生成するベクトル生成手段２０と、特徴ベクトル間の類似度を計算して各文書データを分類する分類手段３０と、文書データの分類結果を出力する出力手段４０とを有する情報検索装置において、ベクトル生成手段２０は、各文書データを各々解析してキーワード及びキーワード間の関係を抽出し、これら両方の出現頻度に基づいて特徴ベクトルを生成する。



1

【特許請求の範囲】

【請求項１】　複数の文書データを格納する文書データベースと、

各々の上記文書データに対し特徴ベクトルを生成するベクトル生成手段と、

上記特徴ベクトル間の類似度を計算して上記各文書データを分類する分類手段と、

上記文書データの分類結果を出力する出力手段とを有する情報検索装置において、

上記ベクトル生成手段は、上記各文書データを各々解析してキーワード及びキーワード間の関係を抽出し、これら両方の出現頻度に基づいて上記特徴ベクトルを生成することを特徴とする情報検索装置。

【請求項２】　複数の文書データを格納する文書データベースと、

検索式を入力する検索式入力手段と、

各々の上記文書データ及び上記検索式に対し特徴ベクトルを生成するベクトル生成手段と、

上記検索式に対する特徴ベクトルと各々の上記文書データに対する特徴ベクトル間の類似度を計算する類似度計算手段と、

類似度の高い上記特徴ベクトルを有する文書データを出力する出力手段とを有する情報検索装置において、

上記ベクトル生成手段は、上記各文書データ及び検索式を各々解析してキーワード及びキーワード間の関係を抽出し、これらの出現頻度に基づいて上記特徴ベクトルを生成することを特徴とする情報検索装置。

【請求項３】　上記ベクトル生成手段は、上記キーワード間の関係として係り受けの関係を用いることを特徴とする請求項１または２に記載の情報検索装置。

【請求項４】　上記ベクトル生成手段は、上記キーワード間の関係としてキーワード間の距離が近いことを用いることを特徴とする請求項１または２に記載の情報検索装置。

【請求項５】　上記ベクトル生成手段は、同一カテゴリに属するキーワード群に含まれるキーワードもしくはそれを含むキーワード間の関係の出現頻度の代わりに、そのカテゴリを代表するキーワードもしくはそれを含むキーワード間の関係の出現頻度としてそれらの出現頻度をそれぞれ加算したものを用いることを特徴とする請求項１から４のいずれかに記載の情報検索装置。

【請求項６】　上記ベクトル生成手段は、キーワード及びキーワード間の関係の出現頻度に対し、利用者が指定する重みづけに基づいて特徴ベクトルを生成することを特徴とする請求項１から５のいずれかに記載の情報検索装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、電子化された文書データの分類・検索に関し、特に文書データを自動的に分類・検索をする情報検索装置に関するものである。

2

【０００２】

【従来の技術】電子化された文書データの分類・検索に関しては、従来、例えば特開平１１－１１０３９５号公報に示されている情報検索装置が提案されている。ここに提案されている情報検索装置においては、類義語の出現頻度をまとめて特徴ベクトルを生成し、特徴ベクトル間の類似度を計算して各文書データを分類する。また、この特開平１１－１１０３９５号公報には、類義語の関係にある複数の単語にそれぞれ重み付けを付けることも提案されている。

【０００３】さらに、例えば特開平１０－１９８６９１号公報においては、類義語の出現頻度をまとめて特徴ベクトルを生成することが開示されているとともに、文書データベース中の隣接した単語対、類義語対を登録しておき、特徴ベクトルの計算に用いることが開示されている。

【０００４】

【発明が解決しようとする課題】このような構成の従来の情報検索装置においては、語句の係り受け等、キーワード間の関係を反映した分類あるいは検索をしていない、そのため、精度の高い分類あるいは検索をすることが出来なかった。

【０００５】この発明は、上述のような課題を解決するためになされたもので、キーワードだけでなくキーワード間の関係をも反映した類似度計算を可能とし、分類あるいは検索の精度を向上することができる情報検索装置を得ることを目的とする。

【０００６】

【課題を解決するための手段】この発明に係る情報検索装置は、複数の文書データを格納する文書データベースと、各々の文書データに対し特徴ベクトルを生成するベクトル生成手段と、特徴ベクトル間の類似度を計算して各文書データを分類する分類手段と、文書データの分類結果を出力する出力手段とを有する情報検索装置において、ベクトル生成手段は、各文書データを各々解析してキーワード及びキーワード間の関係を抽出し、これら両方の出現頻度に基づいて特徴ベクトルを生成する。

【０００７】また、この発明に係る情報検索装置は、複数の文書データを格納する文書データベースと、検索式を入力する検索式入力手段と、各々の文書データ及び検索式に対し特徴ベクトルを生成するベクトル生成手段と、検索式に対する特徴ベクトルと各々の文書データに対する特徴ベクトル間の類似度を計算する類似度計算手段と、類似度の高い特徴ベクトルを有する文書データを出力する出力手段とを有する情報検索装置において、ベクトル生成手段は、各文書データ及び検索式を各々解析してキーワード及びキーワード間の関係を抽出し、これらの出現頻度に基づいて特徴ベクトルを生成する。

【０００８】また、ベクトル生成手段は、キーワード間

の関係として係り受けの関係を用いる。

【０００９】また、ベクトル生成手段は、キーワード間の関係としてキーワード間の距離が近いことを用いる。

【００１０】また、ベクトル生成手段は、同一カテゴリに属するキーワード群に含まれるキーワードもしくはそれを含むキーワード間の関係の出現頻度の代わりに、そのカテゴリを代表するキーワードもしくはそれを含むキーワード間の関係の出現頻度としてそれらの出現頻度をそれぞれ加算したものを用いる。

【００１１】さらに、ベクトル生成手段は、キーワード及びキーワード間の関係の出現頻度に対し、利用者が指定する重みづけに基づいて特徴ベクトルを生成する。

【００１２】

【発明の実施の形態】実施の形態１．図１は、この発明の分類に関する情報検索装置の構成例を示すブロック図である。図において文書データベース１０は複数の文書データを格納する。文書データベース１０に格納された各文書データは少なくともテキストデータを有している。

【００１３】ベクトル生成手段２０は、各文書データに対して特徴ベクトルを生成する。すなわち、各文書データのテキストデータに対して形態素解析などを行い、必要に応じて不要語処理等を行ってキーワードを抽出すると共に、キーワード間の関係を抽出する。

【００１４】次に、Ｎ個の文書からなる文書データベース１０全体から、キーワードＫ個、キーワード間の関係Ｒ個が抽出されたとき、各文書ｉ（１≦ｉ≦Ｎ）の特徴ベクトルＶｉは、たとえば、Ｋ＋Ｒ次元のベクトルで表される。キーワード若しくはキーワード間の関係のインデックスをｊ（１≦ｊ≦Ｋ＋Ｒ）で表すとき、特徴ベクトルＶｉの各次元ｊの成分Ｖｉｊは、たとえば、ｔｆ・ｉｄｆ法によると、次の式で算出できる。

【００１５】Ｖｉｊ＝ＴＦｉｊ＊ｌｏｇ（Ｎ／ＤＦｊ）

【００１６】ここで、ＴＦｉｊは、文書ｉ中において、ｊ成分に対応するキーワード若しくはキーワード間の関係が現れる回数であり、また、ＤＦｊは、文書データベース１０のＮ個の全文書中において、ｊ成分に対応するキーワード若しくはキーワード間の関係が現れる回数である。このようにして、特徴ベクトルが生成される。

【００１７】分類手段３０では、文書間の類似度を計算すると共に、その結果を使って、文書のクラスタリングを行う。文書間の類似度は、たとえば上記のように算出した複数文書の各特徴ベクトル間の角度のコサイン値で計算できる。クラスタリングについては、Ｋ平均法などのクラスタリングアルゴリズムに必要な類似度計算に、上述のように生成した特徴ベクトルを使って計算した類似度を用いることにより、従来のクラスタリングで用いられていたキーワードだけでなく、キーワード間の関係をも反映したクラスタリングが可能となる。

【００１８】分類した結果は出力手段４０により出力することができる。

【００１９】ここで、ベクトル生成手段２０で、各文書データに対して特徴ベクトルを生成する際のキーワード間の関係抽出の具体例としては、構文解析の結果として得られる係り受けの関係やキーワード間の距離の近いものなどが考えられる。

【００２０】まず、係り受けの関係について、たとえば、「ＡがＢをＣする」という文を考える。この文においては、「ＡがＣする」「ＢをＣする」という係り受けの関係が存在する。これらを格まで含めて識別してもよいが、格を無視して、「Ａ→Ｃ」、「Ｂ→Ｃ」あるいは、方向も無視して、「Ａ＆Ｃ」、「Ｂ＆Ｃ」（「Ａ→Ｃ」と「Ｃ→Ａ」を同じと見なす）とすることも考えられる。この場合のキーワード間の関係に係る、前記ＴＦｉｊ、若しくはＤＦｊとしてはこのような係り受けの出現回数を使用することになる。

【００２１】一方、キーワード間の関係として、キーワード間の距離の近いものを用いる場合の距離の具体例としては、たとえばキーワード間の文字数、形態素数、文節数、文数、段落数等が考えられる。この場合も方向を考える場合と考えない場合が存在する。この場合のキーワード間の関係に係る、前記ＴＦｉｊ、若しくはＤＦｊとしては、例えば、この距離がユーザー指定値より小さい場合の出現回数を使用することになる。

【００２２】さらに、同義語関係などのキーワードのカテゴリを反映した分類を行うことも可能である。すなわち、係り受けの関係を例にとって説明すれば、キーワードａ０、ａ１がカテゴリＡに属しており、ａ０、ａ１はｂと係り受けの関係があるとすれば、キーワードａ０、ａ１に関する次元、「ａ０」、「ａ０→ｂ」、「ａ１」、「ａ１→ｂ」を、「Ａ」、「Ａ→ｂ」に、まとめて、特徴ベクトルを生成することも出来る。

【００２３】なお、分類の際の比較対象となる特徴ベクトルの各次元は、共に非ゼロ成分（共起）でなければ類似度には寄与しない。しかしながら、一般に、キーワード間の関係の共起は、単なるキーワードの共起よりも確率的に低くなるため、キーワードに比べて、キーワード間の関係の類似度への寄与が低くなってしまう傾向がある。そこで、キーワード間の関係に関する次元の重みをキーワードの次元よりも大きくすることにより、両者の類似度評価への寄与のバランスを図ることができる。

【００２４】また、ユーザーがキーワードを選定して、そのキーワードの次元やそのキーワードが含まれるキーワード間の関係の次元の重みを大きくして、ユーザーが注目するキーワードを重視した分類を行うことも可能である。

【００２５】実施の形態２．図２は、この発明の他の実施例である、検索に関する情報検索装置の構成例を示すブロック図であり、文書データベース１０、出力手段４０は図１と同じ機能を有する。

【００２６】検索式入力手段５０は、検索条件を検索式（検索式を表現する文章でも良い。）として入力する機能を有し、ベクトル生成手段２０は、文書データベース１０に格納されている全文書について、キーワード及びキーワード間の関係の出現頻度から、例えば、実施の形態１で述べた特徴ベクトル計算式にしたがって特徴ベクトルを生成すると共に、入力した検索式からも、実施の形態１で述べた特徴ベクトル生成方法と同様の方法を用いて、特徴ベクトルを生成する機能を有する。（ただし、検索式から特徴ベクトルを生成する場合は、前記Ｖｉｊのｉは文書ｉを示すものではなく、検索式を示すものとする。この場合、ＴＦｉｊは通常１になる。）

【００２７】検索手段６０は、検索式から生成した特徴ベクトルと、格納された文書データベース１０中の全文書についての特徴ベクトル間の類似度を、実施の形態１の分類の手段で述べた方法と同様な方法で計算し、その結果を使って、文書データベース中の文書の類似度ランキング評価を行う。

【００２８】ランキング付けした結果は出力手段４０により出力することができる。

【００２９】なお、ベクトル生成手段２０では、各文書データに対して特徴ベクトルを生成する際のキーワード間の関係抽出の具体例として、実施の形態１で述べたことと同様に、キーワードの係り付けの関係やキーワード間の距離の近いもの数などを利用することが考えられる。

【００３０】さらに、同義語関係などのキーワードのカテゴリを反映した検索を行うことも可能である。キーワード若しくはキーワード間の関係のカテゴリーへのまとめ方は、実施の形態１の例示と同様である。

【００３１】また、特定のキーワード若しくはキーワード間の関係の次元の重みを大きくすることで、実施の形態１での例示と同様に、キーワード／キーワード間の関係のそれぞれの次元の重みのバランスを図ったり、ユーザーが注目するキーワード若しくはキーワード間の関係を重視した検索を行うことも可能である。

【００３２】

【発明の効果】この発明に係る情報検索装置は、複数の文書データを格納する文書データベースと、各々の文書データに対し特徴ベクトルを生成するベクトル生成手段と、特徴ベクトル間の類似度を計算して各文書データを分類する分類手段と、文書データの分類結果を出力する出力手段とを有する情報検索装置において、ベクトル生成手段は、各文書データを各々解析してキーワード及びキーワード間の関係を抽出し、これら両方の出現頻度に基づいて特徴ベクトルを生成する。そのため、文書データの分類において、各文書データのキーワードだけでなく、キーワード間の関係をも反映した類似度計算が可能となり精度が向上する。

【００３３】また、この発明に係る情報検索装置は、複数の文書データを格納する文書データベースと、検索式を入力する検索式入力手段と、各々の文書データ及び検索式に対し特徴ベクトルを生成するベクトル生成手段と、検索式に対する特徴ベクトルと各々の文書データに対する特徴ベクトル間の類似度を計算する類似度計算手段と、類似度の高い特徴ベクトルを有する文書データを出力する出力手段とを有する情報検索装置において、ベクトル生成手段は、各文書データ及び検索式を各々解析してキーワード及びキーワード間の関係を抽出し、これらの出現頻度に基づいて特徴ベクトルを生成する。そのため、検索式に近い文書データの検索において、検索式および各文書データに出現するキーワードだけでなく、キーワード間の関係をも反映した類似度計算が可能となり検索の精度が向上する。

【００３４】また、ベクトル生成手段は、キーワード間の関係として係り受けの関係を用いる。そのため、文書データの分類や検索式に近い文書データの検索において、係り受けの関係を反映した類似度計算が行われ、キーワードのみを用いた従来の方式に比べ分類や検索の精度が向上する。

【００３５】また、ベクトル生成手段は、キーワード間の関係としてキーワード間の距離が近いことを用いる。そのため、文書データの分類や検索式に近い文書データの検索において、キーワード間の距離を反映した類似度計算が行われ、キーワードのみを用いた従来の方式に比べ分類や検索の精度が向上する。

【００３６】また、ベクトル生成手段は、同一カテゴリに属するキーワード群に含まれるキーワードもしくはそれを含むキーワード間の関係の出現頻度の代わりに、そのカテゴリを代表するキーワードもしくはそれを含むキーワード間の関係の出現頻度としてそれらの出現頻度をそれぞれ加算したものを用いる。そのため、文書データの分類や検索式に近い文書データの検索において、利用者にとって区別することが不要なキーワードあるいはキーワード間の関係をまとめた類似度計算を行うことができる。その結果、高精度の分類・検索の効率化を図ることが可能となる。

【００３７】さらに、ベクトル生成手段は、キーワード及びキーワード間の関係の出現頻度に対し、利用者が指定する重みづけに基づいて特徴ベクトルを生成する。そのため、文書データの分類や検索式に近い文書データの検索において、利用者の意図を反映して、特定のキーワードあるいはキーワード間の関係を重視あるいは軽視した類似度計算を行うことができる。その結果、利用者の意図をより良く反映した形での分類・検索の高精度化が可能となる。

【図面の簡単な説明】

【図１】　この発明の分類に関連する情報検索装置を示すブロック図である。

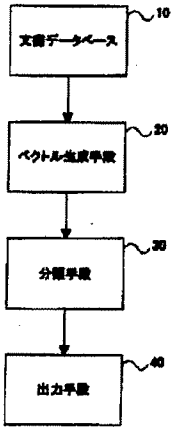
【図２】　この発明の検索に関連する情報検索装置を示

すブロック図である。

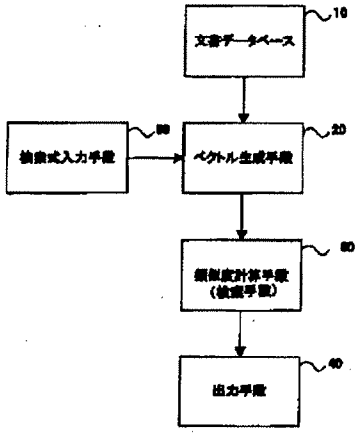
【符号の説明】

１０　文書データベース、２０　ベクトル生成手段、３＊０　分類手段、４０出力手段、５０　検索式入力手段、６０　類似度計算手段（検索手段）。

【図１】



【図２】



フロントページの続き

(72)発明者　小船　隆一
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内

(72)発明者　有田　英一
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内

Ｆターム(参考)　5B075 ND03 NK02 NR12 PP23 PR04 PR06 QM08